CleanAvatar: Artifact-free Gaussian Avatar via Mesh Guidance and Segmentation-driven Augmentation

Jinsong Zhang
Fuzhou University
Fuzhou, China
jinszhang@fzu.edu.cn

I-Chao Shen
The University of Tokyo
Tokyo, Japan
jdilyshen@gmail.com

Jotaro Sakamiya
The University of Tokyo
Tokyo, Japan
jotarosakamiya@g.ecc.u-tokyo.ac.jp

Yuqin Lin
Fuzhou University
Fuzhou, China
linyuqin@fzu.edu.cn

Yu-Kun Lai
Cardiff University
Cardiff, UK
LaiY4@cardiff.ac.uk

Takeo Igarashi*
The University of Tokyo
Tokyo, Japan
takeo@acm.org

Kun Li*
Tianjin University
Tianjin, China
lik@tju.edu.cn

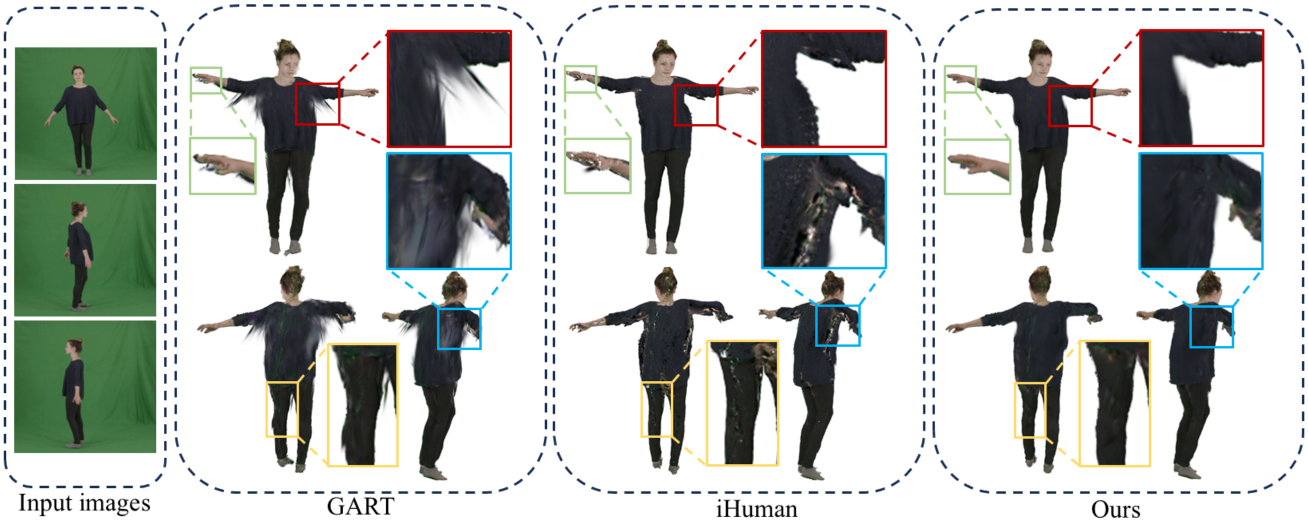


Figure 1: Given only a few input images, our method can reconstruct an animatable human avatar with robust animation results. In contrast, previous work iHuman [29] and GART [20] cannot inpaint invisible regions and produce artifacts during animation.

Abstract

Recent works have adopted 3D Gaussian Splatting (3DGS) to represent animatable avatars. However, these methods require a substantial number of images to train a high-fidelity avatar and often fail to produce photo-realistic images when the driven poses are different from those used during training. To address these challenges, we propose CleanAvatar, a two-stage framework for reconstructing high-quality animatable avatars with a limited number of frames. In the first stage, we optimize a GS avatar and a mesh avatar simultaneously and use the mesh avatar to guide the GS avatar in reducing the invisible regions in training images. Then, we propose a data augmentation method that generates clean rendered images of the GS avatar in unseen poses using a pre-trained segmentation model. Afterwards, we update the GS avatar with both the original and augmented images as training data. Experimental results on three public datasets show that our method can reconstruct high-fidelity avatars from a limited number of input images and produce artifact-free results for unseen poses in comparison to existing methods. The code will be available online.

***Keywords:** Avatar reconstruction, Gaussian splatting, Dual representation.*

*Corresponding authors.

1. Introduction

Human avatar creation plays a central role in immersive applications such as virtual reality, the Metaverse, gaming, and movie production. Early approaches [5, 3] rely on specialized hardware, such as depth sensors or dense multiview camera rigs, which limits their scalability in practical scenarios. This has driven increasing interest in reconstructing animatable avatars directly from sparse monocular video inputs, a setting that is both more practical and significantly more challenging.

Early monocular methods [1, 2] employ mesh-based representations and surface rendering to create personalized avatars. Although efficient, they often produce limited photorealism even for training poses. To address this, subsequent works [47, 31, 15, 8, 14, 12] adopt Neural Radiance Fields (NeRF) [25], which improve fidelity but suffer from heavy computational costs and slow rendering, hindering their usability in interactive applications.

Recent progress in 3D Gaussian Splatting (3DGS) [17] has shown great potential for real-time high-fidelity avatar rendering. Several methods [20, 33, 23] leverage GS as their canonical representation and achieve impressive quality when the target poses closely match the training poses. However, GS parameters such as rotation and scaling often overfit the limited training poses, leading to blurry and distorted artifacts in unseen poses, especially along silhouette boundaries. Hybrid schemes that integrate GS with parametric human templates [29, 37] alleviate overfitting but restrict the expressiveness of GS, thereby lowering rendering quality. Moreover, these approaches struggle to infer invisible regions unseen in the input views. Diffusion-prior-based methods such as HaveFun [43] attempt to inpaint invisible regions and improve animation robustness by leveraging a mesh proxy. While effective to some extent, diffusion priors tend to hallucinate appearances inconsistent with the training images, which is undesirable for applications requiring faithful reproduction.

In this work, we introduce CleanAvatar, a two-stage framework that explicitly addresses both invisible-region completion and boundary artifacts in monocular avatar reconstruction. Our design is motivated by two insights: (1) Mesh avatars provide strong geometric priors that can regularize GS in unobserved regions; (2) High-quality silhouette masks extracted from a segmentation-prior model can be used to generate clean synthetic training data that mitigate boundary artifacts without introducing appearance hallucination.

In the first stage, we jointly optimize a mesh-based avatar and a GS-based avatar in the canonical space. The mesh avatar acts as a geometric teacher, guiding the GS avatar to infer the appearance of invisible regions by enforcing rendering consistency between the two. This mesh-guided supervision significantly improves the completeness and faithfulness of the GS avatar.

In the second stage, we propose CleanMask, a segmentation-driven data augmentation scheme. We observe that modern segmentation models such as Segment Anything Model (SAM) [18], trained on massive clean image corpora, can predict smooth and artifact-free silhouettes even for flawed GS renderings of novel poses. Rather than serving as a generative prior, SAM is exploited here as a mask extractor: we render the avatar under diverse unseen poses, extract reliable silhouettes using SAM, and blend them with the corresponding GS renders to create artifact-free synthetic images. Retraining the GS avatar with both original and cleaned images enhances its robustness to novel poses and suppresses boundary artifacts.

Experimental results on three datasets demonstrate the superior performance of our method across multiple aspects, including the following: 1) high-quality rendering; 2) robustness to novel poses; 3) faithful reproduction of training images; and 4) requiring only a few images. As shown in Figure 1, given three images, our method can reconstruct an animatable human avatar with artifact-free animation results. Our main contribution can be summarized as follows:

- We propose a novel two-stage avatar reconstruction model, dubbed CleanAvatar, which generates high-quality and animation-robust avatar without the need for substantial training images.
- We propose a mesh-guided Gaussian avatar generation method, to obtain an avatar that infers the invisible regions and faithfully matches the training images.
- We propose a data augmentation method to obtain additional artifact-free training images using a pre-trained segmentation model. To the best of our knowledge, this is the first time that segmentation prior is adopted to remove artifacts for data augmentation.

2. Related Work

2.1. Mesh-based Reconstruction

With the human template model, *i.e.*, SMPL [24] and SMPL-X [30], many methods [46, 1, 2, 13, 22] have adopted mesh-based representations with surface rendering to model human avatars. To improve geometric accuracy, these methods typically add displacements to the vertices of the template model and use texture maps to represent color information. For instance, Alldieck *et al.* [2] introduced a visual hull method to optimize the geometry of SMPL using monocular video, leading to a personalized blend shape model. Zhao *et al.* [46] proposed a dynamic surface network to reconstruct pose-dependent geometry and coarse texture, and then refined the rendering with a reference-based neural rendering network. To reduce the number of input images needed, Alldieck *et al.* [1] predicted geometry based on SMPL with vertex displacements directly from monocular images and employed a graph cut-based optimization method using eight frames to compute the texture map. These approaches produce mesh avatars that are robust for animating novel poses. However, due to the fixed topology and limited resolution of both the mesh and texture map, modeling fine-grained appearance details remains a challenge, resulting in suboptimal rendering outcomes.

In this paper, rather than relying solely on a mesh representation, we adopt 3D Gaussian Splatting to achieve more realistic results. Additionally, we introduce a mesh avatar to assist in inpainting the invisible regions of the 3DGS avatar.

2.2. Implicit Function-based Reconstruction

To achieve better rendering results, many methods [31, 40, 44, 8, 4, 9] have adopted Neural Radiance Fields (NeRF) [25] as their canonical representation, and adopt SMPL [24] as the deformation guidance. These methods learn an implicit function, *i.e.*, occupancy, via multilayer perceptron (MLP). With this representation, these methods achieve promising results and can be used to model loose clothes. However, the computational complexity and slow rendering speed of NeRF limit these methods' suitability for real-time applications. While some methods [15, 16] have attempted to reduce training time and improve rendering speed, they still require a large number of images to generate a high-quality avatar.

To address these challenges, we propose CleanAvatar, which can generate high-quality and artifact-free avatar from a limited number of input images.

2.3. Gaussian Splat-based reconstruction

3D Gaussian Splatting (3DGS) [17] is an emerging alternative that offers an efficient approach to 3D scene reconstruction by representing objects through a sparse set of 3D Gaussians, which can achieve high-quality reconstructions with real-time rendering. Many methods [23, 27, 20, 33, 28, 6] introduced GS as their representation to model avatars in the canonical space. These methods produce realistic images when driving poses are similar to the training images, but the rendered results often contain artifacts when the driving poses fall significantly outside the training set. Therefore, these methods require a large number of images with diverse poses to train the avatar. To address these challenges, some methods [29, 39, 37, 45] have integrated mesh representations by binding GS to each mesh face, achieving more robust animation results and reducing the number of required training images. However, mesh representations can limit the flexibility of 3DGS, leading to suboptimal outcomes. Furthermore, GS-based methods, due to their discrete nature, cannot inpaint invisible regions in the training images.

In this paper, we propose a mesh-guided Gaussian avatar generation method that produces a high-quality avatar from a small number of images. Our approach alleviates the limitations of mesh-embedded GS avatar reconstruction methods and can inpaint invisible regions.

3. Preliminaries

3.1. Linear Blend Skinning

To model animatable humans, a popular way is modeling geometry and appearance in a canonical space [33, 20]. Then, linear blend skinning (LBS) [21] based on the human parametric template, *e.g.*, SMPL [24], is used to deform the avatar from the canonical space to the observation space. In practice, we adopt SMPL with $n = 24$ joints as our human template. Then, for a point x_c in the canonical space, we adopt LBS to deform it to the observation space. Mathematically, this can be expressed as:

$$x_o = LBS(x_c; B(\theta)) = \sum_{k=1}^{n} \mathcal{W}(x_c) B_k x_c, \tag{1}$$

where $\mathcal{W}$ denotes the function that queries skinning weights of a given point, which can be formulated by a predefined or learnable skinning weight field.

3.2. 3D Gaussian Splatting (3DGS)

3DGS [17] is a static representation, which models the scene using a set of Gaussian primitives. Consider a Gaussian $\mathcal{G}_i = \{\mu_i, r_i, s_i, o_i, f_i\}$, where μ represents the 3D mean, r denotes the 3D rotation, s corresponds to the scaling factors along the three axes, o is the opacity, and f represents the spherical harmonics coefficients. The 3D covariance is given by:

$$\Sigma_i = r_i \mathrm{diag}(\mathrm{s_i^2}) r_i^{\mathrm{T}}. \tag{2}$$

To project a 3D Gaussian into the 2D image plane, a viewing transformation W and the Jacobian J of the affine approximation of the projective transformation are applied. The 2D covariance $\Sigma^{'}$ is defined as:

$$\Sigma^{'} = JW\Sigma W^{\mathrm{T}} J^{\mathrm{T}}. \tag{3}$$

After that, the color of each pixel can be computed using alpha blending:

$$C = \sum_i o_i^{'} (\Pi_{j=1}^{i-1} (1 - o_j^{'})) c_i, \tag{4}$$

where $o_i^{'}$ is the density contribution weighted o_i by the 2D covariance, and c_i is computed by evaluating view-dependent Spherical harmonics with coefficients f_i.

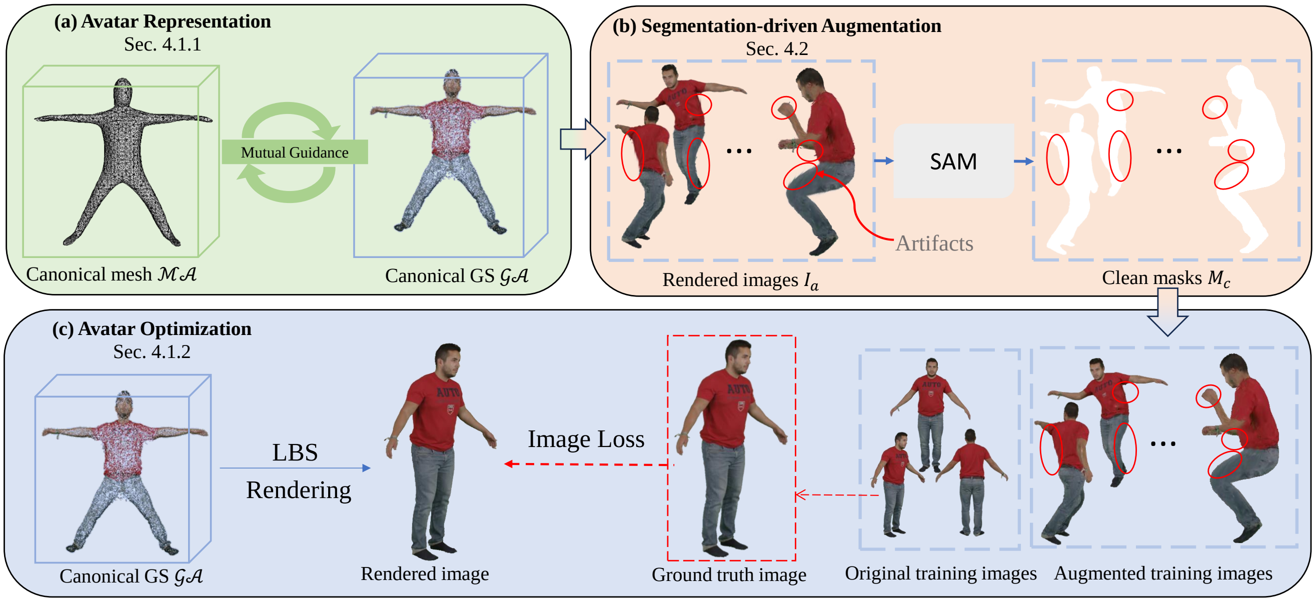


Figure 2: Overview of our method. In the first stage, *i.e.*, (a) DualAvatar, we simultaneously optimize a mesh avatar and a GS avatar in the canonical space. Next, we leverage the GS avatar to render additional images in unseen poses. To remove artifacts in these additional images, we use SAM to predict clean masks and generate artifact-free images using Equation 11 (b). Finally, we train the final GS avatar, $\mathcal{GA}_f$, using the original training images combined with the augmented images in unseen poses (c).

4. Method

Given a set of images $\{I_t\}_{t=1}^{M}$ from a monocular camera, we aim to reconstruct a high-quality Gaussian Splatting (GS) avatar capable of robust animation results. To achieve this, we propose CleanAvatar, a two-stage framework to reconstruct high-quality avatars from monocular images (Figure 2). The most significant differences compared with existing methods are that we adopt mesh representation as a guidance to inpaint invisble regions in training images, and leverage the pretrained segmentation model to augment clean animation results as training data. Inspired by [1], we observe that, by using position-aware functions, the mesh avatar with surface rendering can obtain reasonable results for invisible regions in training images. Therefore, in the first stage, by introducing a mesh avatar to supervise the GS avatar in the canonical space, we first obtain a preliminary avatar that is faithful to the training images and can inpaint invisible regions. However, the animation results often exhibit many artifacts in boundary regions due to limited observations. To obtain additional images in different poses, we find that the pretrained segmentation model can remove artifacts effectively. In the second stage, we first animate the flawed avatar from the first stage to obtain many images in arbitrary poses, and adopt Segment Anything Model (SAM) [18] to remove artifacts in rendered images. Then, we retrain a GS avatar using original training images and additional artifact-free images, resulting in a high-quality avatar with robust animation performance.

4.1. DualAvatar Generation

Previous works [20, 33, 29] ignore the invisible regions in the training images, and thus require a large number of images to train a GS avatar, limiting the practical applications. HaveFun [43] introduces a 3D diffusion prior and mesh representation to inpaint invisible regions, and achieves promising results on avatar reconstruction from few-shot images. However, the diffusion prior can lead to hallucinated appearance, leading to plausible but unfaithful results compared to the input images. We observe that mesh-based avatars, due to surface renderings one-to-one ray-surface correspondence, inherently produce compact and occlusion-aware color estimation. As demonstrated in Figure 3, optimizing texture maps on an SMPL-based mesh yields precise visible-region colors, whereas GS avatars based on volume rendering leak artifacts into unobserved areas. Although mesh rendering alone may lack photorealism, we leverage its geometric discipline to reduce invisible regions via position-aware color prediction, ensuring fidelity to the input. To bridge realism and robustness, we propose DualAvatar, a hybrid framework that synergizes mesh and GS representations. By co-optimizing both avatars in canonical space, we transfer the meshs structured inpainting to the GS avatar, achieving faithful completion of occluded regions without diffusion-induced hallucinations.

4.1.1 Avatar Representation

We adopt 2DGS [11] as the representation of our GS avatar, and utilize Deep Marching Tetrahedra (DMTet) [38] as the representation of our mesh avatar. Both avatars are defined in the canonical space, *i.e.*, "Da" pose, and can be animated using LBS (details can be found in the supplemental document). The GS avatar, denoted as $\mathcal{GA}$, is presented as:

$$\mathcal{GA} = \{\mu_i, r_i, s_i, o_i, f_i\}_{i=1}^{N}, \tag{5}$$

where N is the total number of Gaussians, μ_i represents the Gaussian mean, r_i denotes rotation, s_i indicates scaling factors, o_i is opacity, and f_i refers to spherical harmonics coefficients.

The mesh avatar $\mathcal{MA} = (V, F)$, where V denotes the vertices and F represents the mesh faces, is derived from a tetrahedra representation using DMTet [38]. For each vertex in the tetrahedra grid, we adopt a learnable signed distance field (SDF) to compute the SDF value, enabling the differential extraction of the mesh avatar $\mathcal{MA}$ [38]. To obtain the color of each vertex, we adopt a multilayer perceptron (MLP) with a hash table encoder. Given the position of each vertex v_i in the mesh avatar, its color v_i^c is computed as:

$$v_i^c = g(h(v_i)), \tag{6}$$

where g denotes the MLP layer and h is the hash encoder [15]. This design helps to reduce invisible regions in the mesh avatar, as unsupervised vertices with similar positions are likely to produce similar colors.

4.1.2 Avatar Optimization

With the canonical GS avatar and mesh avatar, we need to deform both avatars from the canonical space to the observation space according to the pose parameters θ estimated from input images I. To obtain the deformed GS avatar $\mathcal{GA}_o$, we take each Gaussian $\mathcal{GA}_i$ as a vertex, and deform it using linear blend skinning (LBS). Because we aim to reconstruct avatars from a limited number of images, we do not model dynamic details using the learnable skinning field. Instead, we directly use the skinning field of SMPL as the predefined skinning prior. Similarly, the mesh avatar is deformed to the observation space, denoted as $\mathcal{MA}_o$.

To optimize both avatars, we render images and minimize the difference between the rendered images and the ground truth images. Gaussian Rasterization [11] is used for rendering the GS avatar and Nvidiffrast [19] is employed for rendering the mesh avatar. Let $\hat{I}(\mathcal{GA}_o)$ and $\hat{I}(\mathcal{MA}_o)$ denote the rendered images of the GS and mesh avatars, respectively. We adopt a combination of an $\mathcal{L}_1$ term and a SSIM term [20] to optimize both avatars:

$$\mathcal{L}_{go} = \mathcal{L}_1(I, \hat{I}(\mathcal{GA}_o)) + \mathcal{L}_{SSIM}(I, \hat{I}(\mathcal{GA}_o)), \tag{7}$$

$$\mathcal{L}_{mo} = \mathcal{L}_1(I, \hat{I}(\mathcal{MA}_o)) + \mathcal{L}_{SSIM}(I, \hat{I}(\mathcal{MA}_o)), \tag{8}$$

where $\mathcal{L}_{go}$ and $\mathcal{L}_{mo}$ represent the image loss terms for the GS avatar and the mesh avatar, respectively.

The GS avatar and the mesh avatar each have their own advantages. On the one hand, the GS avatar benefits from the efficiency of GS rasterization, making it easier to train effectively, whereas the mesh avatar is challenging to optimize due to the lack of explicit geometric information. On the other hand, the mesh avatar excels at inpainting invisible regions, a capability the GS avatar lacks. To harness their complementary strengths, we employ mutual guidance between them.

GS-guided reconstruction for mesh avatar. Reconstructing a high-quality mesh avatar from limited images is challenging due to insufficient geometric supervision. To address this, we introduce a Chamfer Distance loss $\mathcal{L}_{CD}$ between the GS avatar and the mesh avatar in the canonical space, leveraging the geometric structure of the GS avatar to guide the optimization of the mesh avatar.

Mesh-guided refinement for GS avatar. To reduce the invisible regions of the GS avatar, we minimize the difference between the rendered images of the mesh avatar and the GS avatar in the canonical space. However, it is difficult to distinguish the invisible regions for the GS avatar. Therefore, in the first stage, though the low-quality mesh images can affect the quality of the GS avatar in visible regions of the training images, we adopt the $\mathcal{L}_1$ term applied to all pixels in the rendered images, which can be formulated as:

$$\mathcal{L}_{gc} = \mathcal{L}_1(\hat{I}(\mathcal{GA}_o), sg(\hat{I}(\mathcal{MA}_o))), \tag{9}$$

where sg indicates that the gradient is stopped to prevent backpropagation through the rendered mesh images.

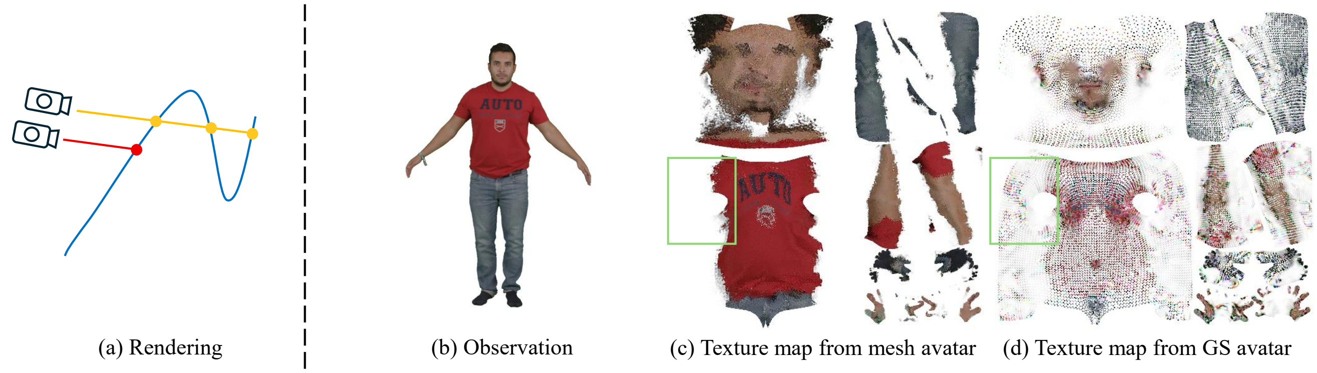
Therefore, the overall loss function in the first stage is:

$$\mathcal{L}_f = \mathcal{L}_{go} + \alpha\mathcal{L}_{mo} + \beta\mathcal{L}_{CD} + \gamma\mathcal{L}_{gc}, \tag{10}$$

where α, β and γ are the weighting factors that balance different loss terms.

4.2. CleanMask for Artifacts Removal

Previous works [20, 33] adopt anisotropy regularizer to obtain better animation results. However, they fail to obtain robust results when driving poses are different from training poses, leading to the need of substantial number of training images with varied poses [42]. Though some approaches [29] combine GS with mesh representations to improve animation results, these methods can limit the flexible capacity of GS, ultimately reducing the quality of rendered results. To obtain robust animation results while


(a) Rendering

(b) Observation

(c) Texture map from mesh avatar

(d) Texture map from GS avatar

Figure 3: The motivation of mesh guidance. (a) Surface rendering (red rays) enforces one-to-one ray-to-surface correspondence, enabling precise texture optimization in visible regions (c). In contrast, volume rendering suffers from alpha-compositing artifacts due to uncontrolled blending along rays, which results in visible discontinuities and unrealistic texture reconstruction in occluded areas (d).

preserving the capacity of GS, we find that universal image semantic segmentation model, *i.e.*, Segment Anything (SAM) [18, 35], can generate clean masks even when the input images contain artifacts, as shown in Figure 2 (b). The primary reason enabling SAM to deliver artifact-free masks is its training on clean, artifact-free image pairs, allowing it to function as a strong prior for generating accurate human masks even in the presence of artifacts. Leveraging this insight, we propose CleanMask, a novel approach to augment training images by animating the avatar generated in the first stage and cleaning the animation results. While some approaches [41] adopt generative models to inpaint invisible regions for 3D reconstruction, introducing additional texture information, our method instead uses a segmentation model to remove artifacts. This avoids introducing any external texture, guaranteeing the final avatar's fidelity to the original training images.

Given the avatar $\mathcal{GA}$ from the first stage, we render additional images in unseen poses. For each rendered image I_a containing artifacts, we first estimate the mask M_c using SAM. Then, the clean image I_c is obtained by applying the output mask as the artifacts cleaner to the original image I_a:

$$I_c = I_a \times M_c. \tag{11}$$

This straightforward process can remove the artifacts in the rendered image, and the clean image I_c can be used as an augmented training image.

By applying the CleanMask method, we generate a new set of training images in previously unseen poses. Using these augmented images and the original training images, a new GS avatar $\mathcal{GA}_f$ can be obtained according to Section 4.1.2. It should be noted that, the augmented images do not introduce additional texture information, unlike diffusion priors used in some few-shot methods [43]. This ensures that the final avatar remains faithful to the original training images. Because the mesh information is stored in the avatar from the first stage and we take the additional images rendered by the GS avatar $\mathcal{GA}$, we do not use the mesh avatar in the second stage, which means that we only adopt $\mathcal{L}_{go}$ as the loss function for the second stage.

4.3. Training details

We initialize the GS (Gaussian splatting) avatar using SMPL body mesh in the "Da" pose, and adopt the densify-and-prune strategy [11] during optimization. In the first stage, we start by optimizing the mesh avatar and GS avatar for $1,500$ iterations. During this process, GS-guided reconstruction $\mathcal{L}_{CD}$ is adopted to obtain a robust mesh avatar. Therefore, α, β, γ are set to $1, 10, 0$. After that, the mesh avatar is fixed. Then, we train the GS avatar with the guidance of the mesh avatar $\mathcal{L}_{gc}$ for another $1,500$ iterations. α, β, γ are set to $0, 0, 1$ during this process. While the mesh avatar is robust for invisible regions, the mesh-rendered images are blurry, which can influence the texture of GS avatar. To alleviate the effect of the blurry mesh-rendered images, we adopt mesh-guided refinement every 10 iterations. For each subject, the whole GS avatar is trained in $3,000$ iterations. In the second stage, the clean mask is obtained by querying the "human", "shirt" and "pants" for images using [35]. The training time for stage-1 is about 6 minutes, data augmentation takes about 1 minute, and stage-2 is about 2 minutes. Overall, the total training time is around 9 minutes on an NVIDIA 3090 GPU. We will make our code available for research purposes.

5. Experiment

5.1. Evaluation Setting

We conduct experiments on avatar reconstruction task from limit monocular video observations on three datasets,

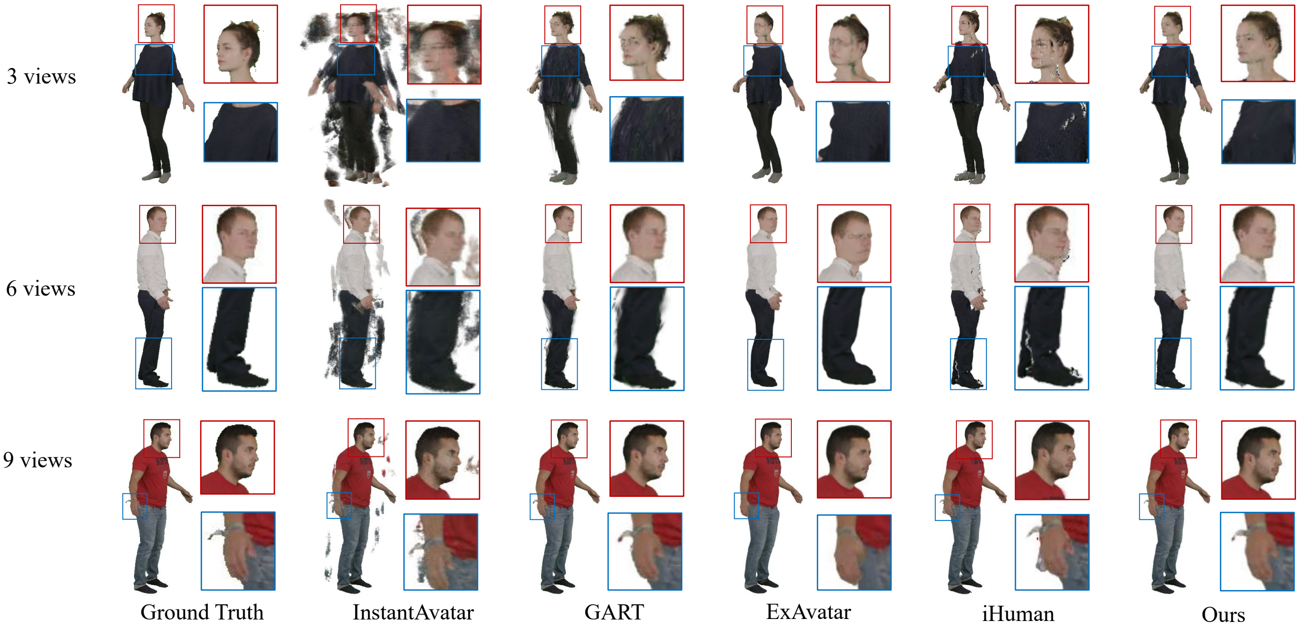


Figure 4: Qualitative comparison with several methods on People-Snapshot [2] test set.

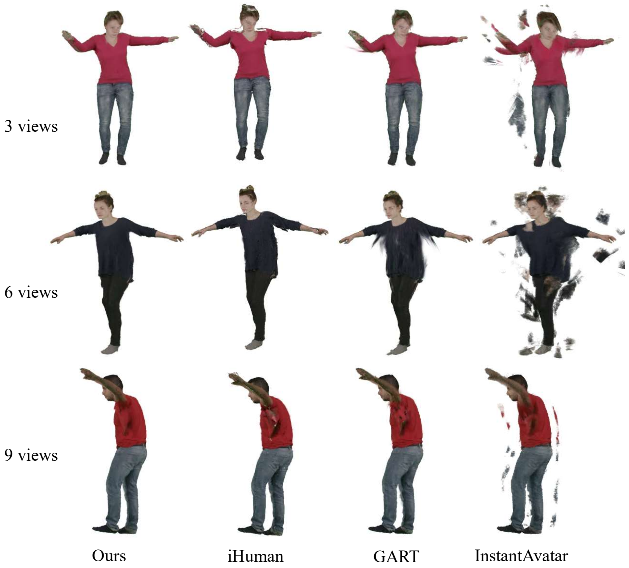


Figure 5: Animation results on People-Snapshot. Please zoom in for details.

including People-Snapshot [2], ZJU-MoCap [32], and FS-XHumans [43].

People-Snapshot [2]. Following previous works [15, 20, 29], we adopt 4 sequences with different identities. This dataset contains monocular videos of a human rotating in front of a camera, resulting in limited pose variations.

ZJU-MoCap [32]. We adopt 6 sequences as the evaluation data for ZJU-MoCap dataset following [20, 33]. This dataset contains multi-view videos of a human in dynamic poses.

FS-XHumans [43]. Introduced by Yang *et al.* [43], this synthetic dataset is designed for avatar reconstruction from few-shot images. For each identity, they take 8 scans to render the training images, and use an A-pose scan to render 24 images from different view as the test data. However, because the test images are not available and it is difficult to evaluate the quality of GS avatar using A-pose images, we render dynamic pose sequences from XHumans as test images.

We uniformly sample different numbers of images as the training data, *i.e.*, 3/6/9 images for People-Snapshot and ZJU-MoCap, and 4/8 images for FS-XHumans. We evaluate the quality of rendering results using commonly used PSNR, LPIPS and SSIM metrics. To better evaluate the rendering results from human perception, we also introduce a recent metric, DreamSim (DS) [7], which is better aligned with human visual perception by tuning large vision models on their collected dataset, as our evaluation metric. Besides, for the People-Snapshot dataset where the test poses are similar with training poses, we adopt FID [36] score to evaluate the quality of animation results by comparing with the original image sequence.

5.2. Comparison

Baselines. We take the recent works GART [20], iHuman [29] and HaveFun [43] as our baselines. GART is a GS-based method, which represents the avatar using GS in

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓
InstantAvatar (3 views)	16.13	0.8261	0.2534	0.2450
GART (3 views)	23.95	0.9375	0.0602	0.0785
ExAvatar (3 views)	22.59	0.9337	0.0632	0.0529
iHuman (3 views)	22.67	0.9271	0.0601	0.0733
Ours (3 views)	**25.25**	**0.9506**	**0.0442**	**0.0408**
InstantAvatar (6 views)	20.52	0.8949	0.1543	0.1582
GART (6 views)	25.52	0.9499	0.0495	0.0505
ExAvatar (6 views)	23.71	0.9422	0.0568	0.0443
iHuman (6 views)	24.43	0.9404	0.0417	0.0473
Ours (6 views)	**26.26**	**0.9572**	**0.0412**	**0.0349**
InstantAvatar (9 views)	23.92	0.9289	0.0966	0.1260
GART (9 views)	26.73	0.9610	0.0400	**0.0215**
ExAvatar (9 views)	24.61	0.9489	0.0536	0.0426
iHuman (9 views)	25.38	0.9486	**0.0327**	0.0276
Ours (9 views)	**27.10**	**0.9637**	0.0373	0.0220

Table 1: Quantitative results on People-Snapshot. We highlight the **first** and second best results for each metric.

	InstantAvatar	GART	iHuman	Ours
3 views	271.3	151.76	130.20	**112.50**
6 views	189.48	122.74	118.23	**105.54**
9 views	149.08	99.93	104.92	**95.79**

Table 2: FID score of animation results on People-Snapshot.

the canonical space. iHuman binds each Gaussian to the triangle face of the human template, *i.e.*, SMPLX, which can reconstruct an animation-robust avatar from few frames of a monocular video. HaveFun is a mesh avatar reconstruction method from few-shot images, which can achieve robust animation results. However, their method needs ground-truth normal maps and depths, and the code of data processing is unavailable, leading to the failure of conducting comparisons on real dataset. To compare more methods, we also take InstantAvatar [15] and ExAvatar [26] as the baseline on People-Snapshot dataset, and compare 3DGS-Avatar [33] on ZJU-MoCap dataset. Besides, we also compare our method with LHM [34], a recent Gaussian avatar reconstruction method, on People-Snapshot dataset, which can be found in the supplemental document.

Quantitative comparison. Table 1 and Table 2 show the quantitative results on People-Snapshot dataset. Given 3/6 images as inputs, our method can achieve best performance on almost all metrics, which demonstrates the effectiveness of our model on avatar reconstruction with limit input images. As the test poses are similar with the training poses, with the number of input images increasing, the invisible regions in the training images become less, but our method still performs better than other methods on most metrics. For LPIPS metric, our method is worse than iHuman, while for DS metric, which is better aligned with human perception than LPIPS, our method performs better than iHuman.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓
3DGS-Avatar (3 views)	27.61	0.9531	0.0545	0.1689
GART (3 views)	28.79	0.9584	0.0513	0.1457
iHuman (3 views)	26.95	0.9484	0.0587	0.1370
Ours (3 views)	**29.56**	**0.9663**	**0.0389**	**0.0982**
3DGS-Avatar (6 views)	28.47	0.9595	0.0441	0.1202
GART (6 views)	29.78	0.9655	0.0412	0.1037
iHuman (6 views)	27.97	0.9529	0.0561	0.1089
Ours (6 views)	**30.18**	**0.9693**	**0.0365**	**0.0822**
3DGS-Avatar (9 views)	28.92	0.9625	0.0390	0.0996
GART (9 views)	30.32	0.9687	0.0371	0.0937
iHuman (9 views)	28.38	0.9562	0.1177	0.1014
Ours (9 views)	**30.65**	**0.9714**	**0.0344**	**0.0778**

Table 3: Quantitative results on ZJU-MoCap.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓
GART (4 views)	21.03	0.9268	0.0713	0.1304
HaveFun (4 views)	21.83	0.9372	**0.0508**	**0.0650**
Ours (4 views)	**21.89**	**0.9386**	0.0533	0.0744
GART (8 views)	21.65	0.9332	0.06 36	0.0899
HaveFun (8 views)	22.03	0.9380	**0.0472**	0.0599
Ours (8 views)	**22.51**	**0.9426**	0.0489	**0.0574**

Table 4: Quantitative results on FS-XHumans.

Besides, our method also achieves best FID score on animation results in unseen poses, which suggests our method can generate robust animation results even using limited input images. Table 3 presents the quantitative results on ZJU-MoCap dataset. Our method achieves best performance on all metrics in different settings, which validates the effectiveness of our method. Table 4 shows the quantitative results on FS-XHumans. We can see that our model achieves better results on pixel-level evaluation (better PSNR and SSIM), but for the 4-views setting, both LPIPS and DS scores are worse than HaveFun. This is because HaveFun has utilized ground-truth normal maps and depth information during training on FS-XHuman, which aids in reconstruction from few-shot images but is unavailable in real-world data.

Qualitative comparison. Figure 4 and Figure 5 show the qualitative results obtained with varying numbers of input images. Previous methods struggle to predict invisible regions, resulting in noticeable artifacts in these areas. For InstantAvatar, there are numerous floating regions when the number of input images is limited. Compared to other methods, our approach effectively infers reasonable textures for invisible regions by leveraging mesh guidance, particularly when the input is limited. As shown in Figure 5, our reconstructed avatar achieves more robust results for unseen poses, particularly in regions invisible during training, such as the armpit area in the third column. Similar conclusions can be drawn from Figure 6, which provides qualitative

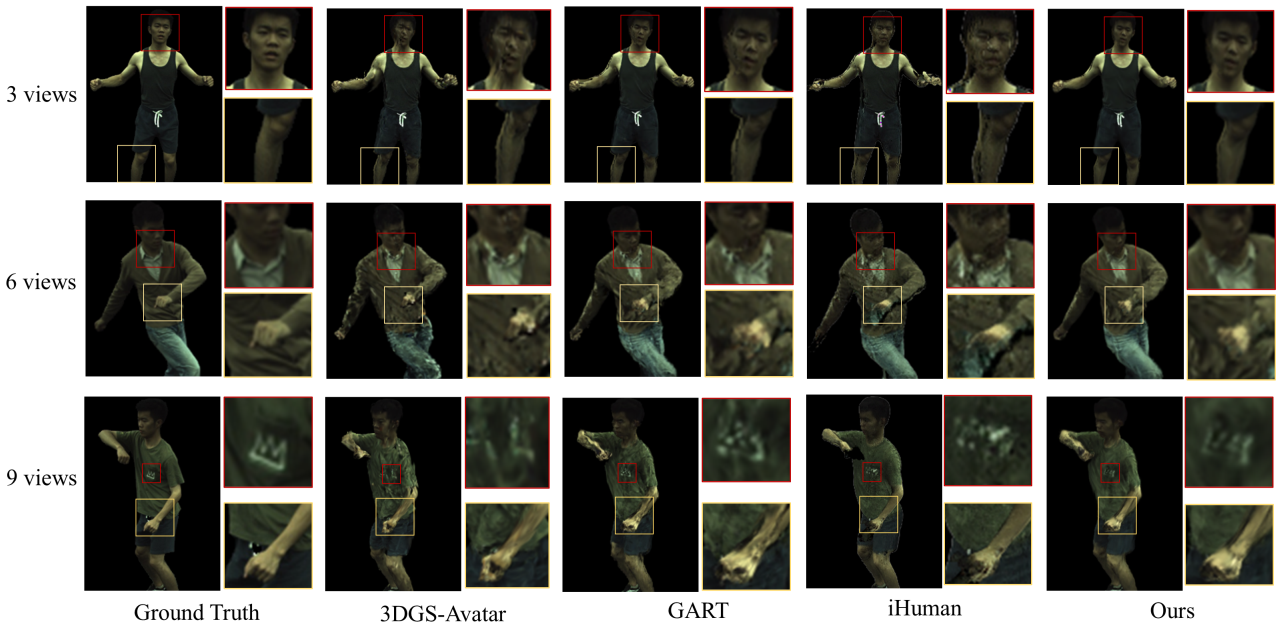


Figure 6: Qualitative comparison on ZJU-MoCap test set.

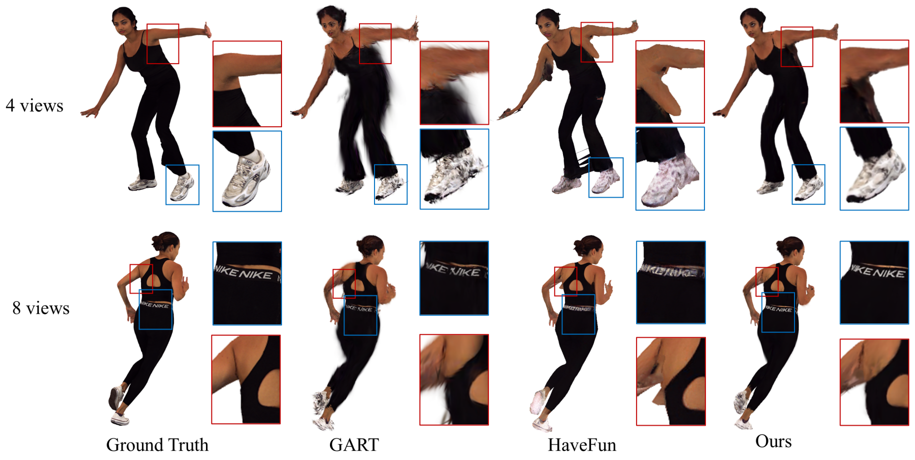


Figure 7: Qualitative comparison on FS-XHumans.

comparisons on the ZJU-MoCap dataset. Our model not only recovers detailed textures in visible regions but also achieves plausible results in invisible regions, where other methods exhibit artifacts (the legs in the first row). For FS-XHumans dataset, as shown in Figure 7, HaveFun generates hallucinated appearances (the color of shoes) due to introducing diffusion priors, and struggles to model detailed textures (words in the second row). Besides, the results are not good in the real data when the ground-truth normal maps and depths are not available (see the supplemental materials). GART can model better textures, but it produces artifacts due to limited training poses. Our method can model detailed texture and achieve robust animation results due to Gaussian splatting representation and CleanMask. More results can be found in the supplemental materials.

In addition, HaveFun [43] requires the ground-truth normal maps and depth maps, and the code for testing on real data is not available. As a result, we are unable to perform comparisons with HaveFun on real-world data. To address this limitation, we extracted the results for the People-Snapshot dataset [2] from their demo video. Figure 8 presents the qualitative comparisons using four input images. The results of HaveFun fail to recover correct color of

https://github.com/TIM2015YXH/HaveFun/issues/1

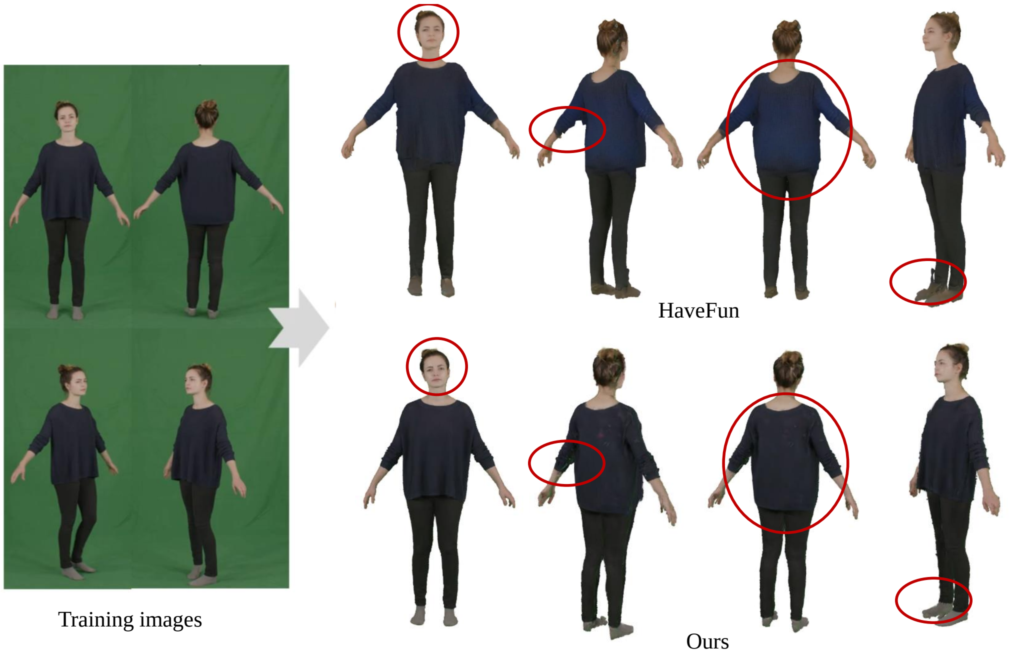


Figure 8: Comparison with HaveFun on People-Snapshot dataset.

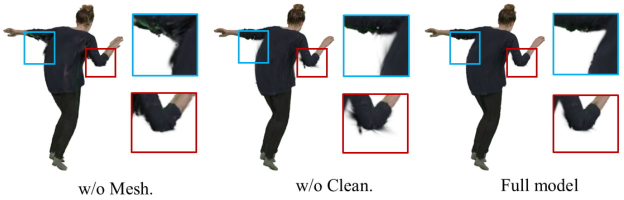


Figure 9: Qualitative comparison of ablation study.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓	FID ↓
w/o Mesh.	26.94	0.9630	0.0374	0.0228	98.43
w/o Clean.	26.80	0.9613	**0.0372**	0.0228	100.78
Ours	**27.10**	**0.9637**	0.0373	**0.0220**	**94.51**

Table 5: Comparison of ablation study on People-Snapshot.

the sweater due to the reliance on the robust but inaccurate diffusion priors. Moreover, the hair and shoes have artifacts due to the absence of ground-truth normal maps and depth maps. In contrast, our method reconstructs a more faithful and visually accurate avatar, benefiting from the robustness of our approach.

5.3. Ablation Study

To validate the design of our method, we design several ablation experiments.

The model without mesh guidance (w/o mesh.). We remove mesh avatar used in our method to validate the effectiveness of mesh guidance.

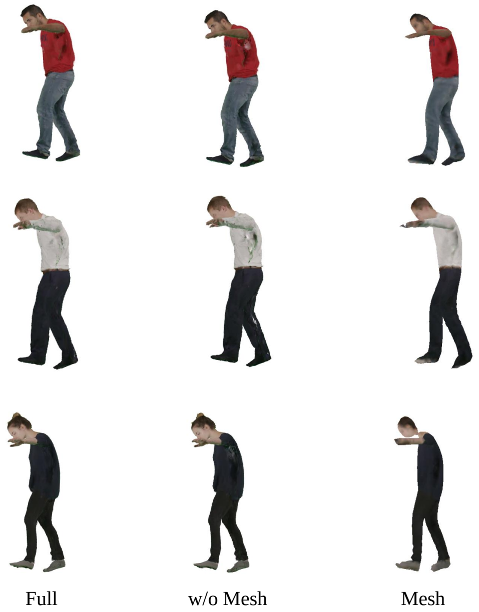


Figure 10: Qualitative results of our mesh avatar. Here, w/o Mesh means the full model with mesh avatar removed, and Mesh means using mesh avatar alone.

The model without CleanMask (w/o clean.). We remove the second stage to validate the segmentation-driven data augmentation method.

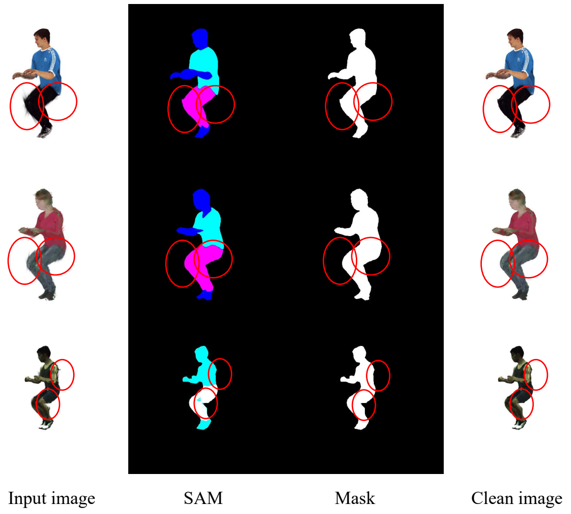


Figure 11: Qualitative results of SAM. Given the input image containing artifacts (the first column), SAM can output smooth and clean masks (the third column), which can be used for artifacts removal (the fourth column).

Table 5 and Figure 9 show the quantitative and qualitative results, respectively. Without mesh guidance, the results have similar scores in SSIM and LPIPS, but have lower performance in other metrics. This is because, as shown in Figure 9, The model without mesh guidance fails to accurately infer and inpaint invisible regions, resulting in incomplete or unrealistic texture reconstruction, particularly in occluded areas. As shown in Figure 9, the model without CleanMask produces many artifacts during animation, leading to poor FID score.

Besides, Figure 10 shows qualitative results of mesh avatars. Though the results are blurry and smooth, the mesh avatar can inpaint missing or invisible regions. This is because 1) the surface rendering avoids ambiguity in volume rendering methods; 2) the position-aware color prediction allows interpolation for invisible regions. However, the inpainting process becomes challenging when the missing regions are extensive, as illustrated in Figure 13.

Analysis of SAM. Our CleanMask approach is motivated by the observation that the pre-trained Segmentation Anything Model (SAM) [18] generates smooth, artifact-free masks even when the input images contain artifacts. Here, we present some results in Figure 11. We select two representative poses containing artifacts for the Gaussian splatting avatar. The first column shows the animation results of the avatar from the first stage, and the second column shows the output segmentation maps using [35] by querying "human", "shirt" and "pants". Though the results are not very precise for different parts, the merged masks are smooth and clean, which can be used to obtain the clean images in the fourth column. This observation is noteworthy and serves as the foundation for artifact removal, *i.e.*, the motivation of our CleanMask.

Results on Different Numbers of Input Images. Our method achieves superior results with a limited number of images. Figure 12 shows the PSNR and DS (DreamSim [7]) scores for different numbers of images on PeopleSnapshot dataset. The results improve as the number of input images increases, but beyond 15 images, the improvements become marginal. Across different numbers of input images, our method consistently outperforms others, with particularly significant advantages when using a small number of input images.

5.4. Failure Cases and Future Work

Although our method can inpaint invisible regions, it is still difficult to achieve high-quality reconstruction from input of only one or two images. Figure 13 shows the results when we adopt two images. For the regions of the side body, the results of mesh avatar are not consistent with the other views, which leads to blurry and unreasonable results. One possible solution is to adopt a human template model with texture mapping, which is easy to inpaint using existing models [10], to represent the mesh avatar. Besides, due to the limited number of input images, we do not model the dynamic deformation, which means that the animation results do not contain dynamic details. This is a common challenge also for other works [29, 43] about avatar reconstruction from monocular videos. A possible direction to overcome this is to model a motion prior for different clothes and poses, which can be used in avatar reconstruction from limited inputs.

6. Conclusion

In this paper, we propose CleanAvatar, a novel two-stage framework for avatar reconstruction that leverages mesh guidance and segmentation-driven augmentation to generate high-quality, artifact-free human avatars from limited images. Experiments on three public datasets validate that our method can not only reduce invisible regions in the training images, but can also achieve artifacts-free animation results.

Acknowledgement

This work was supported in part by National Key R&D Program of China (2023YFC3082100), Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQJC00040), and Natural Science Foundation of Tianjin (24JCYBJC01300).

References

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing

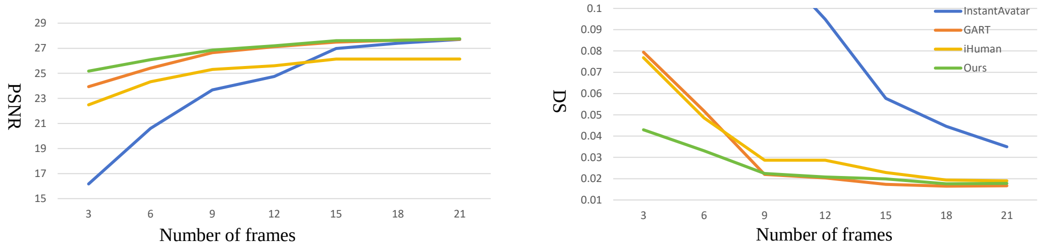


Figure 12: Comparison with other methods on People-Snapshot dataset given different numbers of images.

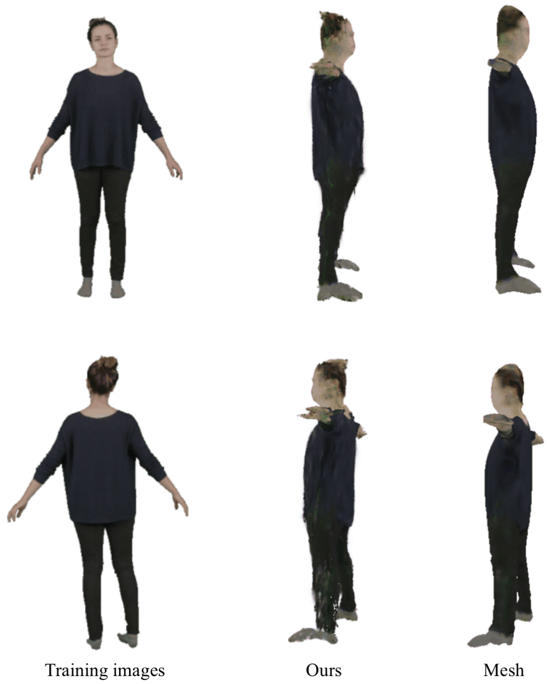


Figure 13: Failure cases for two training images.

from a single RGB camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, jun 2019. 2, 3, 4

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2, 7, 9

[3] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2300–2308, 2015. 2

[4] Y. Chen, X. Wang, X. Chen, Q. Zhang, X. Li, Y. Guo, J. Wang, and F. Wang. Uv volumes for real-time rendering of editable free-view human performance. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16621–16631, 2023. 3

[5] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4):1–13, 2015. 2

[6] X. Deng, Z. Zheng, Y. Zhang, J. Sun, C. Xu, X. Yang, L. Wang, and Y. Liu. Ram-avatar: Real-time photo-realistic avatar from monocular videos with full-body control. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1996–2007, 2024. 3

[7] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 7, 11

[8] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges. Vid2avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 2, 3

[9] C. Guo, T. Jiang, M. Kaufmann, C. Zheng, J. Valentin, J. Song, and O. Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *Eur. Conf. Comput. Vis.*, pages 21–38. Springer, 2025. 3

[10] T. Hu, F. Hong, and Z. Liu. Structldm: Structured latent diffusion for 3d human generation. In *Eur. Conf. Comput. Vis.*, pages 363–381. Springer, 2024. 11

[11] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 5, 6

[12] Z. Huang, S. M. Erfani, S. Lu, and M. Gong. Efficient neural implicit representation for 3d human reconstruction. *Pattern Recognition*, 156:110758, 2024. 2

[13] R. Jena, P. Chaudhari, J. Gee, G. Iyer, S. Choudhary, and B. M. Smith. Mesh strikes back: Fast and efficient human reconstruction from rgb videos. *arXiv preprint arXiv:2303.08808*, 2023. 2

[14] B. Jiang, Y. Hong, H. Bao, and J. Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5605–5615, 2022. 2

[15] T. Jiang, X. Chen, J. Song, and O. Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16922–16932, 2023. 2, 3, 5, 7, 8

[16] Y. Jiang, K. Yao, Z. Su, Z. Shen, H. Luo, and L. Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 595–605, 2023. 3

[17] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 4, 6, 11

[19] S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen, and T. Aila. Modular primitives for high-performance differentiable rendering. *ACM Trans. Graph.*, 39(6), 2020. 5

[20] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis. Gart: Gaussian articulated template models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19876–19887, 2024. 1, 2, 3, 4, 5, 7

[21] J. P. Lewis, M. Cordner, and N. Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 811–818. 2023. 3

[22] X. Li, J. Zhang, Y.-K. Lai, J. Yang, and K. Li. High-quality animatable dynamic garment reconstruction from monocular videos. *IEEE Trans. Circuit Syst. Video Technol.*, 2023. 2

[23] Z. Li, Z. Zheng, L. Wang, and Y. Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19711–19722, 2024. 2, 3

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 3

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[26] G. Moon, T. Shiratori, and S. Saito. Expressive whole-body 3d gaussian avatar. In *Eur. Conf. Comput. Vis.*, pages 19–35. Springer, 2024. 8

[27] A. Moreau, J. Song, H. Dhamo, R. Shaw, Y. Zhou, and E. Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 788–798, 2024. 3

[28] H. Pang, H. Zhu, A. Kortylewski, C. Theobalt, and M. Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1165–1175, 2024. 3

[29] P. Paudel, A. Khanal, A. Chhatkuli, D. P. Paudel, and J. Tandukar. iHuman: Instant animatable digital humans from monocular videos. In *Eur. Conf. Comput. Vis.*, 2024. 1, 2, 3, 4, 5, 7, 11

[30] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10975–10985, 2019. 2

[31] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14314–14323, 2021. 2, 3

[32] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 7

[33] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3dgs-avatar: Animatable avatars via deformable 3D gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5020–5030, 2024. 2, 3, 4, 5, 7, 8

[34] L. Qiu, X. Gu, P. Li, Q. Zuo, W. Shen, J. Zhang, K. Qiu, W. Yuan, G. Chen, Z. Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. In *Int. Conf. Comput. Vis.*, 2025. 8

[35] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 6, 11

[36] M. Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0. 7

[37] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1606–1616, 2024. 2, 3

[38] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Adv. Neural Inform. Process. Syst.*, 34:6087–6101, 2021. 5

[39] J. Wen, X. Zhao, Z. Ren, A. G. Schwing, and S. Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2059–2069, 2024. 3

[40] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16210–16220, 2022. 3

[41] J. Xiang, Y. Guo, L. Hu, B. Guo, Y. Yuan, and J. Zhang. Expressive talking human from single-image with imperfect priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10398–10409, 2025. 6

[42] T. Xie, Z. Zong, Y. Qiu, X. Li, Y. Feng, Y. Yang, and C. Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4389–4398, 2024. 5

[43] X. Yang, X. Chen, D. Gao, S. Wang, X. Han, and B. Wang. Have-fun: Human avatar reconstruction from few-shot unconstrained images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 742–752, 2024. 2, 4, 6, 7, 9, 11

[44] Z. Yu, W. Cheng, X. Liu, W. Wu, and K.-Y. Lin. Monohuman: Animatable human neural field from monocular video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16943–16953, 2023. 3

[45] J. Zhang, X. Li, H. Jia, J. Li, Z. Su, G. Wang, and K. Li. Logavatar: Local gaussian splatting for human avatar mod-

eling from monocular video. *Computer-Aided Design*, page 103973, 2025. 3

[46] H. Zhao, J. Zhang, Y.-K. Lai, Z. Zheng, Y. Xie, Y. Liu, and K. Li. High-fidelity human avatars from a single rgb camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15904–15913, 2022. 2

[47] Y. Zhi, S. Qian, X. Yan, and S. Gao. Dual-space nerf: Learning animatable avatars and scene lighting in separate spaces. In *Int. Conf. on. 3D Vision.*, pages 1–10. IEEE, 2022. 2